

# Uncovering the dynamics of citations of scientific papers

Michael Golosovsky\* and Sorin Solomon

*The Racah Institute of Physics,  
The Hebrew University of Jerusalem,  
91904 Jerusalem, Israel*

(Dated: October 2, 2014)

## Abstract

We demonstrate a comprehensive framework that accounts for citation dynamics of scientific papers and for the age distribution of references. We show that citation dynamics of scientific papers is nonlinear and this nonlinearity has far-reaching consequences, such as diverging citation distributions and runaway papers. We propose a nonlinear stochastic dynamic model of citation dynamics based on link copying/redirection mechanism. The model is fully calibrated by empirical data and does not contain free parameters. This model can be a basis for quantitative probabilistic prediction of citation dynamics of individual papers and of the journal impact factor.

PACS numbers: 01.75.+m, 02.50.Ey, 89.75.Fb, 89.75.Hc

---

\* electronic address: michael.golosovsky@mail.huji.ac.il

## I. INTRODUCTION

The growth mechanism of complex networks is frequently attributed to preferential attachment [1]. While this mechanism accounts for the ubiquity of networks that are scale-free or have heavy-tailed degree distribution, it is too general and does not specifically address evolving network structure. A more realistic scenario of the dynamics of growing networks is provided by the two-step growth models that have been developed in the context of social networks [2, 3], epidemic-like propagation of ideas [4–6], diffusion of innovations [7, 8], and citation dynamics [9]. In the context of citations these models are known as redirection/copying [10], recursive search [11, 12], link copying/referral [13], uniform/preferential attachment [14], and triad formation [15, 16]. Although citation network is specific (it is ordered, directed, acyclic, and does not allow rewiring [17, 18]), it is an excellent example of a growing network since it is well-documented and its dynamics can be reliably traced through long time periods.

We introduce a comprehensive two-step model of a growing citation network. The model is fully calibrated by empirical data and does not contain free parameters. Our measurements revealed an unexpected dynamic nonlinearity that was missing in all previous models. We incorporate this nonlinearity into our framework and come out with a nonlinear stochastic model of citation dynamics. The model predictions are confirmed by the measurements of the age composition of the average reference list on the one hand, and by the statistical distribution of cumulative citations for a large ensemble of papers, on another hand.

Our model can be useful for making probabilistic prediction of citations of scientific papers. This active topic was initiated by Refs. [19–23] who suggested several linear predictive models containing empirical parameters. Statistical uncertainty of these one-step models is too high. We introduce here a much more realistic nonlinear two-step model where all parameters have been calibrated in the independent measurements. The nonlinearity leads to divergent citation dynamics that explains why predicting citation behavior of individual papers is so difficult. Our model can be a basis for the probabilistic forecasting the scientific impact of a paper or of a journal.

## II. STATISTICS OF REFERENCES

### A. Scenario: how an author composes his reference list

Consider a cartoon scenario of an author writing a scientific paper. He reads research journals or media articles, searches the databases, finds the relevant papers and cites some of them in his reference list. Then he studies the reference lists of these preselected papers, picks up relevant references, reads them, cites some of them, and the process continues recursively. We distinguish between the direct references that the author found through media or database search, and indirect references that the author picked up from the reference lists of the preselected papers [9, 13, 14]. Figure 1 shows the corresponding reference network.

The direct and indirect references emerge in another scenario where the author finds each reference independently. Since old references are usually seminal studies, the author's most recent references will probably cite these old papers as well. In our parlance the older references are indirect ones although the author could choose them without knowing that other preselected papers cite them as well.

### B. Age distribution of references

The above scenario yields a very specific age distribution of references. Indeed, consider a reference list of an average paper that comprises  $R_0 = \int_0^\infty R(t_0, t_0 - t) dt$  references where  $t_0$  is the publication year and  $R(t_0, t_0 - t)$  is the number of references that were published in the year  $t_0 - t$ . The latter consist of the direct and indirect references,  $R(t_0, t_0 - t) = R_{dir}(t_0, t_0 - t) + R_{indir}(t_0, t_0 - t)$ . For example, Fig. 2 shows  $R(t)$ ,  $R_{dir}(t)$ ,  $R_{indir}(t)$  for Physics papers published in  $t_0 = 1984$ .

To find  $R(t)$  we make a crude approximation that once the author cites some paper, he can cite any of its references *with equal probability*. An average reference list comprises  $R(t_0, t_0 - \tau)$  preselected papers published in year  $t_0 - \tau$  (where  $\tau < t$ ), each of which bringing in average  $T(\tau)$  indirect references. The fraction of the latter that were published in the year  $t_0 - t$  is  $\frac{R(t_0 - \tau, t_0 - t)}{R_0(t_0 - \tau)}$ . This is equal to  $\frac{R(t_0, t_0 - t + \tau)}{R_0(t_0)}$  since the age composition of the average

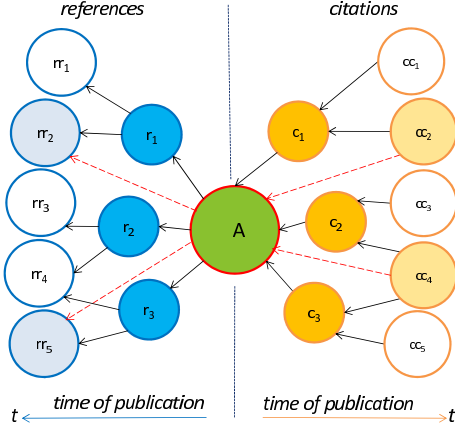


FIG. 1: A fragment of a citation network showing two generations of references and citations of some parent paper A. The circles depict the papers, the solid lines depict direct references (citations), the dashed lines depict indirect references (citations). Each indirect reference (citation) closes a triangle where the parent paper A is a vertex. The papers  $r_1, r_2, r_3$  are direct references of the paper A. The second-generation papers  $rr_1, rr_2, \dots, rr_5$  appear in the reference lists of the first-generation papers  $r_1, r_2, r_3$ . Some of the former ( $rr_2, rr_5$ ) also appear in the reference list of the parent paper A and we call them indirect references. The papers  $c_1, c_2, c_3$  (first generation citing papers) cite paper A directly. The papers  $cc_1, cc_2, \dots, cc_5$  (second generation citing papers) cite the first generation citing papers  $c_1, c_2, c_3$ . Some of the former ( $cc_2, cc_4$ ) also cite the parent paper A and we call them indirect citations.

reference list is fairly independent of the publication year (Fig. 2). Finally, we obtain

$$R(t) = \underbrace{R_{dir}(t)}_{\text{direct}} + \underbrace{\int_0^t R(\tau) \frac{T(\tau)}{R_0} R(t - \tau) d\tau}_{\text{indirect}}. \quad (1)$$

Since all variables now refer to the same publication year  $t_0$ , we can drop  $t_0$  from our notation, in such a way that  $R(t)$  in Eq. 1 denotes  $R(t_0, t_0 - t)$ .

Once we know the functions  $R_{dir}(t)$  and  $T(t)$ , we can solve Eq. 1 and find  $R(t)$ . These functions are the key parameters of the model since they capture the citation habits of an average author. Although these functions could be found by analyzing reference lists of papers, this is not easy since bibliometric databases focus more on citations than on references. However, since there is a duality between citations and references, we can find

$R_{dir}(t)$  and  $T(t)$  by considering citation dynamics of the papers.

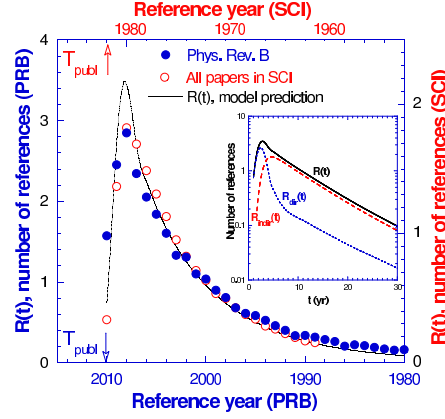


FIG. 2:  $R(t)$ , a number of references in the reference list of a typical paper, that were published in a certain year  $t$ . Blue circles show an average over all papers published in Physical Review B in 2010 (excluding overviews). Red circles show corresponding data for all papers covered by Science Citation Index that were published in 1982 (Ref.[24]). The arrows show publication year of the parent papers (2010 for the PRB papers and 1982 for the SCI papers). Both dependences are almost identical (the difference in  $y$ -scales is due to the fact that the average length of the reference list in 2010 was  $R_0 = 35$  while  $R_0 = 21.5$  in 1982). This identity shows that the age composition of the average reference list does not depend on the publication year of the parent paper. The solid blue line shows model prediction based on Eq. 1 with  $R_0 = 30$  and  $T = 6.6e^{-0.64(t-1)}$  where  $t = 1$  is the publication year. The inset shows model prediction for  $R(t)$ ,  $R_{dir}(t)$ ,  $R_{indir}(t)$ - the total, direct, and indirect references, correspondingly (see Eq. 1).

### III. CITATION DYNAMICS- A MEAN-FIELD MODEL

#### A. Reference-citation duality

Since one paper's citation is another paper's reference, the reference and citation networks are dual (Fig. 1). Consequently, the age distribution of references  $R(t)$  for the papers published in one year (diachronous or retrospective citation distribution [24, 25]) (Fig. 2) is very similar to  $M(t)$  (Fig. 3), the mean citation rate of the papers published in one year (synchronous or prospective citation distribution).

In what follows we analyze consequences of this duality and how it can be used to measure relevant parameters in Eq. 1. Indeed, consider a set of all  $N_0(t_0)$  papers in a certain research field that were published in year  $t_0$ . The mean number of citations that a paper from this set garners in the  $t$ -th year after publication is  $M(t_0, t_0 + t)$ . Since the majority of citing papers belong to the same research field, the total number of citations garnered by these  $N_0(t_0)$  papers in the year  $t_0 + t$  shall be equal to the total number of the references in the reference lists of the papers published in the year  $t_0 + t$ ,

$$N_0(t_0)M(t_0, t_0 + t) = N_0(t_0 + t)R(t_0 + t, t_0) \quad (2)$$

Equation 2 relates the total number of citations and references for the papers belonging to the same research field but published in different years. To find corresponding relation for the papers published in the same year, we shall take into account that both the number of publications  $N_0$  and the reference list length  $R_0$  grow exponentially with time,  $N_0(t_0) \propto e^{\alpha t_0}$ ,  $R_0(t_0) \propto e^{\beta t_0}$ . We substitute these exponential dependences into Eq. 2, notice that  $\frac{R(t_0+t, t_0)}{R_0(t_0+t)} = \frac{R(t_0, t_0-t)}{R_0(t_0)}$ , and find the mathematical expression for the reference-citation duality,

$$M(t_0, t_0 + t) = e^{(\alpha+\beta)t} R(t_0, t_0 - t). \quad (3)$$

## B. A mean-field model of citation dynamics

The substitution of Eq. 3 into Eq. 1 yields dynamic equation for the mean citation rate of the papers published in one year

$$M(t) = \underbrace{M_{dir}(t)}_{\text{direct}} + \underbrace{\int_0^t M(t-\tau) \frac{T(t-\tau)}{R_0} M(\tau) d\tau}_{\text{indirect}} \quad (4)$$

where  $M_{dir} = R_{dir}(t)e^{(\alpha+\beta)t}$  and  $T(\tau)$  has been replaced by  $T(t-\tau)$  using the properties of the convolution. Equation 4 tells us that an average paper published in some year  $t_0$  has  $M(\tau)$  first-generation citations published in year  $t_0 + \tau$ , each of which generating  $M(t-\tau)$  second-generation citations in some later year  $t_0+t$ . The probability that a second-generation citation induces (indirect) citation of the parent paper is  $T(t-\tau)/R_0$  where  $R_0$  is the average length of the reference list of the papers published in the year  $t_0$ .

To solve Eq. 4 we have to determine functions  $M_{dir}(t)$  and  $T(t)$ . In fact, we used Eq. 4 to find  $T(t)$ . To this end we measured  $M(t)$  and  $M_{dir}(t)$  (see next section), substituted

these functions into Eq. 4, and found an exponential kernel  $T = T_0 e^{-\gamma(t-1)}$  where  $T_0 = 6.6$ ,  $\gamma = 0.64 \text{ yr}^{-1}$ , and the publication year corresponds to  $t = 1$ . (These numbers mean that  $\sim 35\%$  references in the average reference list of a paper are direct and  $\sim 65\%$  are indirect).

To find age distribution of references  $R(t)$  we calculated  $R_{dir}(t) = M_{dir}(t)e^{-(\alpha+\beta)t}$  where for Physics papers  $\alpha = 0.046, \beta = 0.02$  (as found in our independent measurements). We substituted  $R_0 = 35$ , and the functions  $R_{dir}(t), T(t)$  found above into Eq. 1 and solved it to find  $R(t)$ . Figure 2 shows that the model prediction  $R(t)$  fits perfectly well our measurements. We conclude that the mean-field model (Eqs.1,2,4) faithfully accounts for the average age composition of the reference list and for the mean citation dynamics of scientific papers. In what follows we use this model to infer citation dynamics of individual papers.

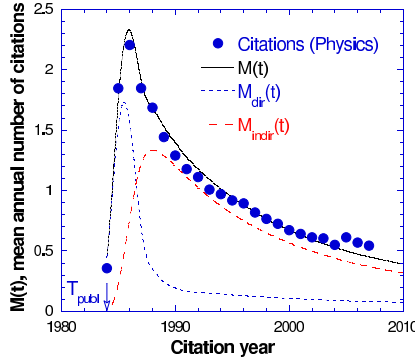


FIG. 3:  $M(t)$ , the mean annual number of citations garnered by a typical Physics paper published in 1984 (blue circles). The black line shows model prediction based on Eq. 4 with  $R_0 = 21.5$  and  $T = 6.6e^{-0.64(t-1)}$  where  $t = 1$  is the publication year. The dashed lines show the direct  $M_{dir}(t)$  and indirect  $M_{indir}(t)$  citations.  $M_{dir}(t)$  shoots up 1-2 years after publication and then sharply decays, while  $M_{indir}(t)$  achieves its maximum after 3-4 years and then slowly decays.

## IV. CITATION DYNAMICS OF INDIVIDUAL PAPERS

### A. Linear stochastic model

To infer equation describing citation dynamics of individual papers we consider  $\Delta k^A$ , the number of citations garnered by some paper  $A$  during a short time interval from  $t$  to  $t + \Delta t$ .

The cumulative number of citations of this paper is  $k^A(t) = \sum_0^t \Delta k^A$ . We assume that  $\Delta k^A$  is a discrete random variable that follows a time-inhomogeneous Poisson process [26] with the rate  $\lambda^A(t)$ . This latent citation rate [27] consists of the direct and indirect contributions,  $\lambda^A(t) = \lambda_{dir}^A(t) + \lambda_{indir}^A$ .

To infer dynamic equation for  $\lambda^A$  we note that for the set of papers published in one year, the average rate of direct citations is  $M_{dir}(t) = \overline{\lambda_{dir}^A(t)}$ , the average rate of total citations is  $M(t) = \overline{\lambda^A(t)}$ , and  $M(\tau)d\tau = \overline{\Delta k^A(\tau)}$  where  $d\tau = 1$ . We substitute these equalities into Eq. 4, replace integral by sum, dispense with the averaging, and obtain

$$\lambda^A(t) = \underbrace{\lambda_{dir}^A(t)}_{\text{direct}} + \underbrace{\sum_{\tau=0}^t M(t-\tau) \frac{T(t-\tau)}{R_0}}_{\text{indirect}} \Delta k^A \quad (5)$$

This discrete stochastic equation is consistent with Eq. 4. Our initial assumption (to be revised soon) is that the functions  $M(t-\tau)$  and  $T(t-\tau)$ , determined from our studies of mean-field citation dynamics, govern citation dynamics of individual papers as well.

## B. Measurements and comparison to the model

To verify Eq. 5 empirically we need to measure the direct and indirect citations separately. To this end we chose 37 representative research papers that were published in the Physical Review B in one year, analyzed their first- and second-generation citing papers (Fig. 1), identified the direct and indirect citations and measured their dynamics. As an aggregate measure of the paper's individuality we took  $k_\infty$ , the long-time limit of cumulative citations.

### 1. Direct citations

We found (not shown here) that the direct citation rate of a paper can be represented as

$$\lambda_{dir}^A(t) = p^A m(t) \quad (6)$$

where  $p^A$  is the numerical parameter and the function  $m(t)$  is the same for all papers published in one year, whereas  $\int_0^\infty m(t)dt = 1$ . Figure 3 shows that  $m(t)$  grows immediately after publication of the paper, achieves its maximum after  $\sim 2$  years and slowly decays thereafter. The long tail of  $m(t)$  is a mathematical expression of delayed recognition ("sleeping beauty" [31]) phenomenon.



The parameter  $p^A$  is the long-time limit of the number of direct citations and it is a proxy to the so-called "fitness" [13, 28] which shall depend on the scientific quality of the parent paper, the journal where it was published, popularity of the research field, etc. On the one hand,  $p^A$  can be estimated *a priori* from the initial citation rate of the paper. [Indeed, shortly after publication  $\frac{dk^A}{dt}|_{t=1} \approx \lambda_{dir}|_{t=1} = p^A m|_{t=1}$ .] On another hand, since the solution of Eqs. 5,6 yields  $k_\infty^A \propto p^A$ , this relation can serve as an *a posteriori* estimate of  $p^A$ . Our measurements (not shown here) yield a sublinear dependence

$$p^A = 0.72(k_\infty^A + k_0)^{0.8} \quad (7)$$

where small parameter  $k_0 \approx 1$  accounts for the fact that previously uncited papers have some probability to be cited in future.

## 2. Indirect citations

We found that Eq. 5 with the kernel  $\frac{T_0}{R_0} M(t - \tau) e^{-\gamma(t-\tau)}$  fits the dynamics of indirect citations of individual papers only if we allow for  $T_0$  and  $M$  to depend on the number of previous citations  $k$ . The reason for this surprising  $k$ -dependence is that new citations modify the very structure of citation network associated with the cited paper, this modification being most pronounced for highly-cited papers.

Indeed, consider two generations of citing papers associated with a parent paper. Obviously, the number of the first-generation citations  $k$  is equal to the number of the first-generation citing papers. However, the numbers of second generation citations and citing papers can differ. We denote by  $M$  and  $N$ , correspondingly, the long-time limits of the number of second-generation citations and citing papers per one first-generation citing paper, and introduce  $r = M/N$ , an average number of the first-generation citing papers cited by a second-generation citing paper. Figure 5 shows that  $r$  increases with  $k$  following empirical dependence  $r = 1 + 0.11 \log k + 0.033(\log k)^2$ . The inset shows that this growth is associated with the  $M(k)$  dependence (this means that the citation network is assortative) while  $N$  is almost independent on  $k$ . Figure 4 demonstrates that  $r$  measures the average number of paths leading from the parent paper to a second-generation citing paper. For a low-cited parent paper  $r = 1$ , indicating that it is connected to each of its second-generation descendant by a single path. For a highly-cited parent paper  $r > 1$  indicating that it is connected

to some of its second-generation descendants by multiple paths.

We found that the parameter  $T_0$  in Eq. 5 is also  $k$ -dependent (not shown here). Therefore, we merge all  $k$ -dependent parameters together and introduce  $P_0 = rT_0/R_0$ , the probability that a second-generation citing paper cites the parent paper (indirectly). Our measurements revealed that  $P_0$  increases nonlinearly with  $j$ , the number of citation paths connecting the parent paper with its second-generation descendant, as it is schematically shown in Fig. 4c. In particular, we found quadratic dependence  $P_0 \propto j^2$  indicating constructive interference between multiple paths. Since the number of multiple paths increases with  $k$ , this translates into empirical dependence  $P_0(k) = 0.16[1 + 3(r - 1)]$  (while in the absence of multipath interference one would obtain  $P_0(k) = 0.16r$ ) The  $P_0(k)$  dependence stems from the  $r(k)$  dependence and it has loc

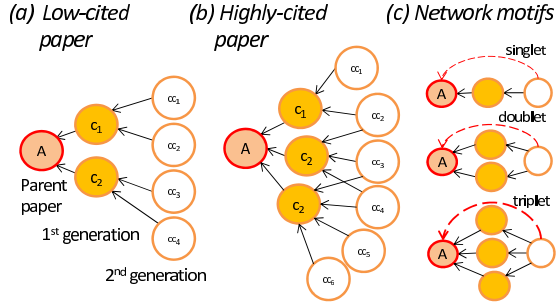


FIG. 4: (a) Two generations of the papers citing a low-cited parent paper. Each second-generation paper cites only one first-generation paper and it is connected to the parent paper by a single path. The numbers of the second-generation citations  $M$  and citing papers  $N$  are equal,  $r = M/N = 1$ . (b) Two generations of the papers citing a highly-cited parent paper. Each second-generation paper can cite several first-generation papers and it can be connected to the parent paper by multiple paths ( $cc_1, cc_5, cc_6$  are connected to the parent paper by a single path;  $cc_2, cc_3, cc_4$  are connected to the parent paper by double paths). The numbers of second-generation citations and citing papers are not equal,  $r = M/N = 1.5$ . (c) Network motifs. Solid lines show direct citations, dashed line show indirect citations. The probability of indirect citation progressively increases from singlet to doublet to triplet as  $\approx 1 : 4 : 9$  indicating constructive multipath interference.

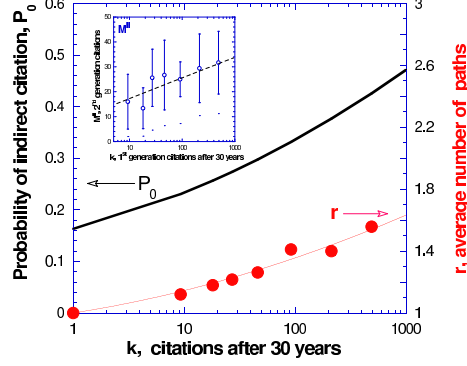


FIG. 5: Microscopic parameters of citation dynamics and their dependence on the number of cumulative citations  $k$ . The solid circles show  $r = M/N$ , the ratio of the numbers of the second-generation citations  $M$  to second-generation citing papers  $N$  (analysis of 108 PRB papers published in 1984).  $r$  characterizes the average number of paths connecting a second-generation citing paper to the parent paper.  $r$  increases with  $k$  following empirical dependence  $r = 1 + 0.11 \log k + 0.033(\log k)^2$  (red solid line). The inset shows that  $M^{II}$ , the number of the second-generation citations per one first-generation citing paper, slowly increases with  $k$ . The solid black line shows  $P_0$ , the probability of indirect citation of the parent paper by a second-generation citing paper. It follows functional dependence  $P_0 = a[1 + b(r - 1)]$  derived from our model where parameters  $a = 0.16$  and  $b = 3$  were found from the microscopic measurements.

## V. NONLINEAR MODEL OF CITATION DYNAMICS

### A. Dynamic equation

To introduce nonlinearity into Eq. 5 we replaced the kernel  $\frac{T_0}{R_0} M(t - \tau) e^{-\gamma(t-\tau)}$  by  $P_0(k) N(t - \tau) e^{-\gamma(t-\tau)}$ . Here,  $N(t)$  is the average number of the second generation citing papers per one first-generation citing paper (fan-out coefficient), and  $P_0$  is the probability that a second-generation citing paper cites the parent paper (indirectly). The novelty here is the  $P_0(k)$  dependence which is shown in Fig. 5. We introduce this kernel into Eq. 5, plug there Eq. 6, and obtain our key result- nonlinear stochastic dynamic equation for the latent

citation rate of a paper A-

$$\lambda^A(t) = p^A m(t) + \sum_{\tau=0}^t P_0(k^A) e^{-\gamma(t-\tau)} N(t-\tau) \Delta k^A(\tau) \quad (8)$$

The empirical functions  $m(t)$ ,  $N(t)$ , and  $P_0(k)$  are shown in the Figs. 3,5, correspondingly. Equation 8 is a nonlinear first-order discrete stochastic differential equation with the initial condition set by  $p^A$ . This equation expresses  $\lambda^A(t)$ , the latent citation rate of the paper A at time  $t$ , through past citations of the same paper,  $\Delta k^A(\tau)$  and  $k^A = \sum \Delta k^A(\tau)$ . The probability distribution of actual citations at time  $t$  is given by the Poisson distribution,  $P(\Delta k) = \frac{(\lambda^A)^{\Delta k}}{(\Delta k)!} e^{-\lambda^A}$ .

## B. Stochastic simulation

To verify Eq. 8 we performed stochastic numerical simulation imitating citation dynamics of a set of 40195 Physics papers published in 1984. Figure 6 shows the cumulative citation distributions for this set over the time span of 25 years. We wish to imitate these distributions using Eq. 8. This requires that the statistical distribution of initial conditions ("fitness"  $p^A$ ) for the actual and "simulated" papers be the same. We estimated  $p^A$  for each paper using Eq. 7 and assuming  $k_\infty \approx k(t = 25)$ . The inset in Fig. 6 shows corresponding statistical distribution of  $p^A$ .

We run stochastic simulation based on Eq. 8 with this distribution of  $p^A$  and empirical functions  $m(t)$ ,  $N(t)$ ,  $P_0(k)$  shown in Figs. 3,5, correspondingly. Figure 6 shows excellent agreement between the simulated and measured cumulative citation distributions. Moreover, our simulation accounts fairly well for such intricate characteristics of citation dynamics as stochastic variability, temporal autocorrelation, and the dynamics of uncited papers (not shown here). We present here our measurements with Physics papers while we obtained very similar results with the Mathematics and Economics papers as well.

## C. Analysis of the model

To have more insight into citation dynamics of scientific papers we consider continuous approximation of Eq. 8. Without loss of generality we disregard stochasticity, consider  $k$  as a continuous variable, and approximate the kernel by the exponential,  $P_0(k)N(t-\tau)e^{-\gamma(t-\tau)} \approx$

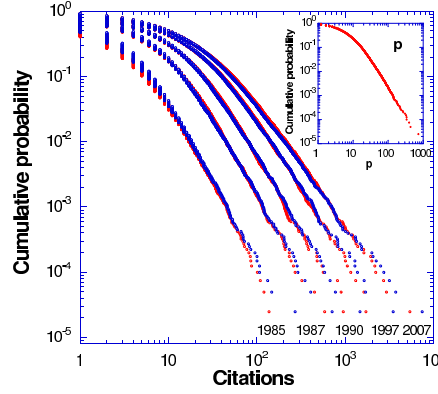


FIG. 6: Cumulative citation distributions for 40,195 Physics papers published in 1984. Red symbols stay for measurements, blue symbols stay for the results of a stochastic simulation based on Poisson process with the rate given by Eq. 8 and the probability  $P_0$  shown in Fig. 5. The inset shows fitness  $p^A$  estimated from Eq. 7.

$q(k)e^{-\gamma'(t-\tau)}$  where  $q \approx 1.26P_0(k)$  and  $\gamma' = \gamma + 0.08$ . [The rationale for this approximation is the fact that  $e^{-\gamma(t-\tau)}$  with  $\gamma = 0.64 \text{ yr}^{-1}$  has much stronger time dependence than  $N(t-\tau)$ . The latter is captured by the term  $0.08 \text{ yr}^{-1}$ ]. We replace the sum in Eq. 8 by the integral, drop index  $A$  and arrive at

$$\frac{dk}{dt} = pm(t) + \int_0^t qe^{-\gamma'(t-\tau)} \frac{dk}{d\tau} d\tau \quad (9)$$

Equation 9 appears in the context of Bellman-Harris branching (cascade) processes [30]. It is well-known in the population dynamics where it describes the age-dependent birth-death process with immigration [29] where direct and indirect citations are analogs of immigration and reproduction, correspondingly, and  $q/\gamma'$  is the reproduction number.

Dynamic behavior described by Eq. 9 results from the interplay between the positive feedback rate characterized by the factor  $q$  and the rate of obsolescence characterized by the parameter  $\gamma'$ . The latter shall be compared to the average paper longevity (citation lifetime),  $\tau_0$ , that we define empirically using a crude exponential approximation  $k(t) = k_\infty[1 - e^{-(t-\Delta)/\tau_0}]$ , where  $\Delta$  characterizes delayed recognition. In the limit  $\gamma'\tau_0 \gg 1$  Eq. 9 reduces to the first-order autoregressive model of citation dynamics [26]

$$\frac{dk}{dt} \approx pm(t) + \frac{q}{\gamma'} \left( \frac{dk}{dt} \right)_{t-1/\gamma'} \quad (10)$$

In the opposite limit,  $\gamma'\tau_0 \ll 1$ , Eq. 9 reduces to the models of Refs. [10, 14]

$$\frac{dk}{dt} \approx pm(t) + qk \quad (11)$$

The latter is nothing else but the Bass equation for diffusion of innovations [7, 8] in an infinite market. Citations correspond to adopters, direct citations correspond to innovators, and indirect citations correspond to imitators. The connection to the Bass model is not occasional since each paper can be considered as a new product whose penetration to the market of ideas is gauged by the number of citations. The novelty here is the nonlinear  $q(k)$  dependence. In the context of diffusion of innovations the nonlinear coefficient of imitation  $q(k)$  is not unexpected. This would indicate increased probability of adoption of a new product if several neighbors in the network already adopted it. To the best of our knowledge, such possibility didn't deserve much attention.

## VI. CONSEQUENCES OF NONLINEARITY: RUNAWAYS

To analyze consequences of the nonlinearity we note that since  $q(k)$  dependence is weak we can integrate Eq. 9 over time assuming constant  $q$ . This yields

$$k(t) \approx p \int_0^t \left[ m(t') + q \int_0^{t'} m(\tau) e^{-(\gamma' - q)(t' - \tau)} d\tau \right] dt' \quad (12)$$

The first term in the square brackets corresponds to direct citations, the second term stays for indirect citations. Each direct citation induces a cascade of indirect citations that propagates in time if  $\gamma' - q < 0$  and decays if  $\gamma' - q > 0$ . In the latter case  $k$  comes to saturation,  $k_\infty \rightarrow p\gamma' / (\gamma' - q)$  (ordinary papers) while in the former case  $k$  grows exponentially (seminal papers). Since  $k$  grows with time, an ordinary paper which by pure chance garnered excessive number of citations, can become a seminal paper.

Although the  $q(k)$  dependence is weak, it is important since it enters in the exponent. This results in a "winner takes all" instability [12, 34–37]. To analyze how this instability develops with  $k$  we again consider the paper longevity (citation lifetime)  $\tau_0$ . The latter is determined by the exponent  $\gamma' - q$ , and to a lesser extent, by the function  $m(t)$ . Equation 12 suggests that  $\tau_0 \propto 1/(\gamma' - q)$ . Since  $q$  increases with  $k$ , the inverse  $\tau_0(q)$  dependence means that with increasing number of citations,  $\tau_0$  increases and diverges upon approaching

the branching (tipping) point  $q(k) = \gamma'$  [13]. This means that each new citation extends a paper's lifetime [38]

Figure 7 demonstrates that the citation lifetime  $\tau_0$  indeed increases with increasing  $k$  and diverges when  $k > 600$ , in such a way that the papers with more than 600-1000 citations exhibit runaway behavior - their citation career does not saturate even after 25 years. This complements the famous parable "rich get richer" by "rich live longer".

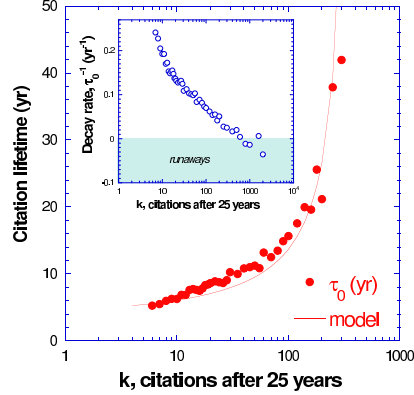


FIG. 7: The paper longevity (citation lifetime),  $\tau_0$ , and the corresponding decay rate  $\tau_0^{-1}$ , versus  $k$ , the number of citations after 25 years. To reduce fluctuations the data were binned.  $\tau_0$  increases with increasing  $k$ , in such a way that highly-cited papers show runaway behavior (diverging  $\tau_0$ , negative decay rate). The solid line shows a crude approximation,  $\tau_0 \propto \frac{1}{\gamma' - q(k)}$ , suggested by Eq. 12. Here,  $\gamma' = 0.72 \text{ yr}^{-1}$  and  $q = 1.26P_0(k)$  where  $P_0(k)$  is shown in Fig. 5.

## VII. SUMMARY

We developed a nonlinear stochastic model of citation dynamics of scientific papers and validated this model by measurements. The underlying scenario is as follows. We assume that the author of a new scientific paper finds relevant papers from the media or journals and cites them. Then he studies the reference lists of these preselected papers, picks up some references, cites them as well, and continues this process recursively. We add here a new ingredient: if some paper is cited by several preselected papers, the author chooses it with higher probability than that cited by only one preselected paper.

This new ingredient, combined with the assortativity of the citation network, introduces dynamic nonlinearity. The account of this nonlinearity is crucial for predicting future citation behavior of the papers. Our nonlinear dynamic model can serve as a basis for probabilistic forecasting of citation dynamics of a paper or a group of papers (journal impact factor).

## Acknowledgments

We are grateful to S. Redner, A. Scharnhorst, L. Muchnik, and D. Shapiro for fruitful discussions, we appreciate instructive correspondence with M. Simkin. We acknowledge financial support of the EU COST Action TD1210.

- 
- [1] R. Albert and A.L. Barabasi, *Statistical mechanics of complex networks*, Reviews of Modern Physics, 74 (2002), pp. 47–97.
  - [2] M.O. Jackson and B.W. Rogers, *Meeting strangers and friends of friends: How random are social networks?*, American Economic Review, 97 (2007), pp. 890–915.
  - [3] D.M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles,, *Winners don't take all: Characterizing the competition for links on the web* , P.N.A.S., 99 (2002), pp. 5207–5211.
  - [4] W. Goffman and V.A. Newill, *Generalization of the epidemic theory. Application to transmission of ideas* , Nature, 204 (1964), pp. 225–228.
  - [5] E. Bruckner, W. Ebeling, and A. Scharnhorst, *The application of evolution models in scientometrics* , Scientometrics, 18 (1990), pp. 21–41.
  - [6] L.M.A. Bettencourt, A. Cintron-Arias, D. I. Kaiser, and C. Castillo-Chavez, *The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models*, Physica a-Statistical Mechanics and Its Applications, 364 (2006), pp. 513–536.
  - [7] F.M. Bass, *New product growth for model consumer durables*, Management Science Series a-Theory, 15 (1969), pp. 215–227.
  - [8] V. B. Lal, Karmeshu, and S. Kaicker, *Modeling Innovation Diffusion with Distributed Time-Lag* , Technological Forecasting and Social Change, 34 (1988), pp. 103–113.
  - [9] N. K. Vitanov and M. R. Ausloos, in *Models of Science Dynamics*, ed. by A.Scharnhorst, K.Borner, and P. van den Besselaar, Springer, Berlin, (2012), pp. 69–126.



- [10] P. L. Krapivsky and S. Redner, *Network growth by copying*, Physical Review E, 71 (2005), p. 036118.
- [11] A. Vazquez, *Disordered networks generated by recursive searches*, Europhysics Letters, 54 (2001), pp. 430–435.
- [12] A. Vazquez, *Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations*, Physical Review E, 67 (2003), p. 056104.
- [13] M. V. Simkin and V. P. Roychowdhury, *A mathematical theory of citing*, Journal of the American Society for Information Science and Technology, 58 (2007), pp. 1661–1673.
- [14] G. J. Peterson, S. Presse, and K. A. Dill, *Nonuniversal power law scaling in the probability distribution of scientific citations*, P.N.A.S., 107 (2010), pp. 16023–16027.
- [15] Z.-X. Wu and P. Holme, *Modeling scientific-citation patterns and other triangle-rich acyclic networks*, Physical Review E, 80 (2009), p. 037101.
- [16] F.-X. Ren, H.-W. Shen, and X.-Q. Cheng, *Modeling the clustering in citation networks*, Physica A, 391 (2012), pp. 3533–3539.
- [17] J. P. Bagrow and D. Brockmann, *Natural Emergence of Clusters and Bursts in Network Evolution*, Phys. Rev. X 3 (2013), p. 021016.
- [18] B. Karrer and M.E.J. Newman, *Random Acyclic Networks*, Phys. Rev. Lett. 128 (2009), p.128701.
- [19] O. Penner, R. K. Pan, A. M. Petersen, K. Kaski, and S. Fortunato, *The case for caution in predicting scientists’ future impact*, Physics Today, 66 (2013), pp. 8–9.
- [20] D.E. Acuna, S. Allesina, and K.P. Kording, *Predicting scientific success*, Nature, 489 (2012), pp. 201–202.
- [21] A. Mazloumian, *Predicting Scholars’ Scientific Impact*, Plos One, 7 (2012), p. e49246.
- [22] A.L. Barabasi, C.M. Song, and D.S. Wang, *Handful of papers dominates citation*, Nature, 491 (2012), pp. 40–40.
- [23] B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones, *Atypical Combinations and Scientific Impact*, Science, 342 (2013), pp. 468–472.
- [24] H. Nakamoto, *Synchronous and diachronous citation distributions*, in Informetrics 87/88, Belgium : Diepenbeek, pp. 157–163 (1988), ed. by L. Egghe and R. Rousseau.
- [25] W. Glanzel, *Towards a model for diachronous and synchronous citation analyses*, Scientometrics, 60 (2004), pp. 511–522.

- [26] M. Golosovsky and S. Solomon, *Stochastic Dynamical Model of a Growing Citation Network Based on a Self-Exciting Point Process*. *Phys. Rev. Lett.*, 109 (2012), p. 098701.
- [27] Q.L. Burrell, *Predicting future citation behavior*, *Journal of the American Society for Information Science and Technology*, 54 (2003), pp. 372–378.
- [28] G. Bianconi and A.L. Barabasi, *Bose–Einstein condensation in complex networks*, *Physical Review Letters*, 86 (2001), pp. 5632–5635.
- [29] W. Ebeling, A. Engel, and V.G. Mazenko, *Modeling of selection processes with age-dependent birth and death rates*, *BioSystems* 19(1986), pp. 213–221.
- [30] T. E. Harris, *The Theory of Branching Processes*, Springer–Verlag, Berlin, 2002.
- [31] A.F.J. van Raan, *Sleeping beauties in science*, *Scientometrics*, 59 (2004), pp. 467–472.
- [32] J.L. Iribarren and E. Moro, *Branching dynamics of viral information spreading*, *Physical Review E*, 84 (2011), p. 046116.
- [33] M.E.J. Newman, *The first-mover advantage in scientific publication*, *EPL*, 86 (2009), p. 68001
- [34] P.L. Krapivsky and S. Redner, *Organization of growing random networks*, *Physical Review E*, 63 (2001), p. 066123.
- [35] S. Zhou and R.J. Mondragon, *Accurately modeling the internet topology*, *Physical Review E*, 70 (2004), p. 066108.
- [36] P.L. Krapivsky and D. Krioukov, *Scale-free networks as preasymptotic regimes of superlinear preferential attachment*, *Physical Review E*, 78 (2008), p. 026114.
- [37] D.S. Wang, C.M. Song, and A.L. Barabasi, *Quantifying Long-Term Scientific Impact*, *Science*, 342 (2013), pp. 127–132.
- [38] Quite opposite to the situation described in the famous Balzac’s novel ”La peau de chagrin”.